ISSN: 2278-4632
Vol-13, Issue-11, November 2023

MACHINE LEARNING- BASED AUTOMATIC SOCIAL SENTIMENT CLASSIFICATION

M Ramchander[1], Akash Swamy[2]

[1]Assistant Professor, Department of MCA, Chaitanya Bharathi Institute of Technology (A), Gandipet, Hyderabad, Telangana State, India

[2]MCA Student, Chaitanya Bharathi Institute of Technology (A), Gandipet, Hyderabad, Telangana State, India

ABSTRACT: Our society has entered a new information era as a result of the tremendous development in information sharing on social media. During the COVID-19 epidemic, microblogging services like Twitter were very popular. We created an automated system for extracting positive, negative, and neutral emotions from tweets and classifying them further using machine-learning (ML) approaches. The created framework may aid in understanding our society's emotions amid major occurrences such as the COVID-19 epidemic. Our methodology is unique in that it combines a lexicon-based technique for tweet sentiment analysis and tagging with supervised machine learning methods for tweet categorization. We assessed the hybrid framework using a variety of metrics, including precision, accuracy, recall, and F1 score. According to our findings, the majority of attitudes are either favorable (38.5%) or neutral (34.7%). Furthermore, with an accuracy of 83%, the long short-term memory (LSTM) neural network has been chosen as the framework's preferred ML approach. The assessment findings show that our hybrid methodology has the ability to automatically identify huge quantities of tweets, such as those on COVID-19, based on societal emotions.

Keywords *– COVID-19, Coronavirus tweets, hybrid framework, sentiment analysis, text classification, tweet classification, Twitter.*

1. INTRODUCTION

People's social media postings reveal their worry and sorrow as a result of the widespread COVID-19 outbreak. The widespread infection inflated social media updates such as tweets, messages, and postings. Importantly, in times of crisis, user-generated data on social media may be a valuable source of information. People extensively utilized social media and microblogging platforms like Facebook and Twitter to communicate their ideas, opinions, and responses. Twitter is the third-largest online social networking site among all social networking platforms. The examination of COVID-19 tweets is particularly useful since user tweets represent our society's ideas and feelings throughout the epidemic. The global spread of the Coronavirus elicited a broad

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

spectrum of feelings and views. The COVID-19 pandemic, by definition, has produced widespread uncertainty and dread. People from many countries reacted differently on social networking platforms. The shift in feelings during pandemic periods generated mental disorders in the form of fear, worry, and a variety of other horrible symptoms; the COVID-19 pandemic has contributed to exposing urban inhabitants' vulnerabilities and offers a substantial public health hazard. Tweets with phrases like "updates on confirmed cases," "COVID-19-related fatality," "early indicators of the epidemic," "economic damage," and "preventive measures" suggest worry and dread on microblogging sites. Furthermore, public opinions on COVID-19-related news on microblogging sites have the ability to spread disparate emotions.

The availability of massive amounts of social media data allows for sentiment analysis [3]. Due to the unstructured and noisy nature of the data, analyzing such a massive volume of information is time-consuming [3]. As a result, it is critical to create automated approaches for analyzing and categorizing tweets that reflect societal emotions. To automate sentiment analysis, machine-learning (ML) algorithms may be utilized. The research [6] focused on a single deep-learning (DL)-based strategy for tweet categorization, while our architecture incorporates many ML techniques. Thus, our research contributes to a better knowledge of which ML algorithms work well and which do not for tweet categorization. Furthermore, previous work, such as [1]-[3], concentrated solely on the sentiment analysis task, whereas we investigate a broader scope of the sentiment analysis chain by automating the classification of COVID-19 tweets using a hybrid framework that combines lexicon-based tweet sentiment analysis and labelling with ML techniques for tweet classification.

2. LITERATURE REVIEW

Social media analysis with AI: Sentiment analysis techniques for the analysis of Twitter COVID-19 data:

Recently, there has been an epidemic known as COVID-19 (corona virus) producing acute respiratory syndrome, which was initially seen in China and is now a pandemic. Social media plays an important part in the present situation of the globe being shut up, which leads to social imbalance among individuals. Suicide attempts were reported in the news like leaves. In this chapter, we want to provide a sentiment analysis on covid-19 of people's reactions to choices made by the government or local authorities through Twitter. We present a method for automatically assessing tweets and classifying them as favorable, negative, or neutral. The precision, quantization, and prediction of the sets may be accomplished by combining automata with NLP (natural language processing). Classification might be pattern-based or NLTK-based (Natural language toolkit). The categorized findings are

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

then saved in structures that may be iterated on until the visualization is requested.

Word frequency and sentiment analysis of Twitter messages during Coronavirus pandemic:

The Coronavirus epidemic has taken the globe, as well as social media, by storm. As public knowledge of the disease grew, so did the number of messages, films, and postings recognizing its existence. Twitter had a similar impact, with the number of postings relating to coronavirus increasing at an unprecedented pace in a very short period of time. This research includes a statistical analysis of Twitter posts on this illness that have been posted since January 2020. There have been two sorts of empirical investigations conducted. The first is based on word frequency, while the second is based on the moods of individual tweet messages. Examining the word frequency might help you identify patterns or trends in the terms used on the site. At this important point, this would also reflect on the psyche of Twitter users. The power law distribution was used to represent the frequencies of unigrams, bigrams, and trigrams. The findings were confirmed using the Sum of Square Error (SSE), R2, and Root Mean Square Error (RMSE) (RMSE). This model's goodness of fit is supported by high R2 values and low SSE and RMSE values. Sentiment analysis has been performed to better understand the current sentiments of Twitter users. The corpus included tweets from the general public as well as WHO. The data revealed that the bulk of tweets were positive in polarity, with just roughly 15% being negative.

Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic:

During a crisis, social media data may be a valuable source of information. User-generated communications provide us a glimpse into people's brains during such moments, revealing their emotions and viewpoints. Because of the high number of such signals, a large-scale examination of population-wide trends is now feasible. In this research, we examine the emotion of Twitter communications (tweets) gathered during the first months of the COVID-19 outbreak in Europe. This is done using a neural network and multilingual text embeddings for sentiment analysis. We categorize the findings according to their country of origin and connect their temporal evolution with events in those nations. This enables us to investigate the impact of the scenario on people's emotions. We show, for example, that lockdown announcements are associated with a drop in mood in virtually all examined nations, which quickly rebounds.

Sentiment analysis of Twitter data:

We investigate sentiment analysis using Twitter data. This study makes the following contributions: (1) We add POS-specific prior

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

polarity characteristics. (2) We investigate the usage of a tree kernel to eliminate the requirement for time-consuming feature engineering. The novel features (when combined with previously suggested features) and the tree kernel perform similarly, outperforming the state-of-the-art baseline.

Cross-cultural polarity and emotion detection using sentiment analysis and deep learning—A case study on COVID-19:

How various cultures react and behave in the face of a crisis is reflected in a society's norms and political will to deal with the circumstance. Events, societal pressure, or the necessity of the hour often force choices that do not reflect the desire of the country. While some may be delighted, others may be resentful. Coronavirus (COVID-19) elicited a range of reactions from countries in response to the choices made by their individual governments. Over the last several months, social media has been inundated with messages expressing both favorable and negative feelings about COVID-19, pandemic, lockdown, and hashtags. Despite their near proximity, several neighboring nations responded differently to one another. Denmark and Sweden, for example, despite their numerous similarities, took opposing positions on the choice made by their respective administrations. Nonetheless, their country's backing was almost universal, in contrast to neighboring South Asian nations where citizens expressed concern and animosity. The goal of this research is to examine how individuals from various cultures reacted to the new Coronavirus and how they felt about the following steps made by various governments. Deep long short-term memory (LSTM) models used to estimate sentiment polarity and emotions from extracted tweets have been trained on the sentiment140 dataset to reach state-of-the-art accuracy. The usage of emoticons demonstrated a new and original method of evaluating supervised deep learning models on Twitter messages.

3. METHODOLOGY

The availability of massive amounts of social media data allows for sentiment analysis. Due to the unstructured and noisy nature of the data, analyzing such a massive volume of information is time-consuming. As a result, it is critical to create automated approaches for analyzing and categorizing tweets that reflect societal emotions. To automate sentiment analysis, machine-learning (ML) algorithms may be utilized. The research relied on a single deep-learning (DL)-based strategy for tweet categorization, while our platform incorporates many ML techniques. Thus, our research contributes to a better knowledge of which ML algorithms work well and which do not for tweet categorization. Furthermore, previous work concentrated solely on the sentiment analysis task, whereas we investigate a broader scope of the sentiment analysis chain by automating the classification of COVID-19 tweets using a hybrid framework that combines lexicon-based tweet sentiment analysis

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

and labelling with ML techniques for tweet classification.

Disadvantages:

1. Due to the unstructured and noisy nature of the data, analysing such a vast volume of information is time-consuming.
2. Reduced classification accuracy

To extract the sentiments used to label the tweets, we use the valence-aware dictionary and sentiment reasoner (VADER) lexicon-based approach. To predict attitudes for unique unlabeled test datasets, these tagged tweets are fed into a supervised ML algorithm such as Gaussian Nave Bayes (GNB), multinomial Nave Bayes (MLNB), logistic regression (LR), decision tree (DT), random forest (RF), and Long Short-Term Memory (LSTM). Our innovative hybrid method combines a natural language processing (NLP) lexicon-based strategy with a supervised ML technique to accomplish our goal of autonomous sentiment categorization. To broaden the scope of our study, we also employed a DL-based LSTM neural network.

Advantages:

1. Our hybrid system has the capability of automatically classifying massive quantities of tweets.
2. The possibility for high-speed automated categorization of social attitudes relating to COVID-19 on Twitter.

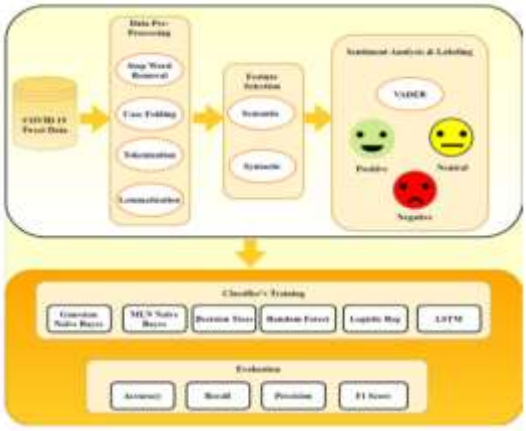

Fig.2: System architecture

MODULES:

To carry out the aforementioned project, we created the modules listed below.

- Data exploration: we will put data into the system using this module.
- Processing: we will read data for processing using this module.
- Splitting data into train and test: Using this module, data will be separated into train and test models.
- Making the model Logistic Regression - Random Forest - AdaBoost - SGD Classifier - KNN - Decision Tree - Multinomial Naive Bayes - SVM - Gaussian Naive Bayes - MLP - Gradient Boosting - Cat Boost - Voting Classifier - LR + RF + SVC - LSTM - RNN - CNN.

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

- User registration and login: Using this module will result in registration and login.
- Using this module will provide input for prediction.
- Prediction: final predicted shown

4. IMPLEMENTATION

ALGORITHMS:

Logistic Regression: Logistic regression is a Machine Learning classification technique that predicts the likelihood of certain classes based on specified dependent variables. In summary, the logistic regression model computes the logistic of the outcome by adding the input characteristics (in most situations, there is a bias component).

Random Forest: A Random Forest Method is a supervised machine learning algorithm that is widely used in Machine Learning for Classification and Regression issues. We know that a forest is made up of many trees, and the more trees there are, the more vigorous the forest is.

AdaBoost: The AdaBoost algorithm, short for Adaptive Boosting, is a Boosting approach used in Machine Learning as an Ensemble Method. Adaptive Boosting is so named because the weights are reassigned to each instance, with larger weights applied to mistakenly categorized instances.

SGD Classifier: Stochastic Gradient Descent (SGD) is a basic yet effective optimization approach for determining the values of function parameters/coefficients that minimize a cost function. In other words, it is used in the discriminative learning of linear classifiers using convex loss functions such as SVM and Logistic regression.

KNN: KNN stands for K-Nearest Neighbors Algorithm. The k-nearest neighbors method, often known as KNN or k-NN, is a non-parametric, supervised learning classifier that employs proximity to classify or predict the grouping of a single data point.

DT: A decision tree is a non-parametric supervised learning technique that may be used for classification and regression applications. It has a tree structure that is hierarchical and consists of a root node, branches, internal nodes, and leaf nodes.

Multinomial Naïve Bayes: The Multinomial Naive Bayes method is a common Bayesian learning strategy in Natural Language Processing (NLP). Using the Bayes theorem, the software estimates the tag of a text, such as an email or a newspaper piece. It computes the probability of each tag for a given sample and returns the tag with the highest chance.

SVM: Support Vector Machine (SVM) is a supervised machine learning technique that may be used for classification and regression. Though

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

we call them regression issues, they are best suited for categorization. The SVM algorithm's goal is to identify a hyperplane in an N-dimensional space that clearly classifies the input points.

Gaussian Naive Bayes: A generative model, Naive Bayes. (Gaussian) Naive Bayes is based on the assumption that each class has a Gaussian distribution. The distinction between QDA and (Gaussian) Naive Bayes is that Naive Bayes assumes feature independence, hence the covariance matrices are diagonal.

MLP: MLPClassifier is an abbreviation for Multi-layer Perceptron Classifier, which links to a Neural Network. Unlike other classification methods such as Support Vectors or Naive Bayes Classifier, MLP Classifier does classification using an underlying Neural Network.

Gradient Boosting: A sort of machine learning boosting is gradient boosting. It is based on the assumption that the best next model, when merged with past models, minimizes the total prediction error. The main concept is to define the desired outcomes for this next model in order to reduce error.

Cat Boost: Cat Boost is a gradient boosting technique for decision trees. It was created by Yandex researchers and engineers and is used for search, recommendation systems, personal assistants, self-driving vehicles, weather prediction, and a variety of other activities at Yandex and other firms such as CERN, Cloudflare, and Careem taxi.

Voting classifier: A voting classifier is a machine learning estimator that trains numerous base models or estimators and predicts based on the results of each base estimator. Aggregating criteria may be coupled voting decisions for each estimator output.

LSTM: LSTM is an abbreviation for Long-Short Term Memory. In terms of memory, LSTM is a sort of recurrent neural network that outperforms standard recurrent neural networks. LSTMs perform far better when it comes to learning specific patterns.

RNN: Recurrent neural networks (RNNs) are the cutting-edge algorithm for sequential data, and they are employed in Apple's Siri and Google's voice search. It is the first algorithm to recall its input thanks to its internal memory, making it ideal for machine learning issues involving sequential data.

CNN: A CNN is a kind of network architecture for deep learning algorithms that is primarily utilized for image recognition and pixel data processing jobs. There are different forms of neural networks in deep learning, but CNNs are the network design of choice for identifying and recognizing things.

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

6. CONCLUSION

We created a unique hybrid system for sentiment analysis in the COVID-19 subject area that combines a lexical method for tweet sentiment analysis and labelling with a DL technique for tweet classification. To automatically categorize social emotions on Twitter, we retrieved positive, negative, and neutral sentiments by labelling COVID-19-related tweets based on their associated feelings using the VADER lexicon approach. We employed several ML and DL algorithms for the classification challenge. With an accuracy of 83% in classification tests, LSTM surpassed all other approaches. When compared to the VADER approach, the trained ML classifier obtained a processing speedup of nearly one order of magnitude. As a consequence, our findings indicated the possibility for high-speed automated categorization of societal emotions connected to COVID-19 on Twitter, which might influence public health PR efforts. To further increase and confirm the high model accuracy level, one intriguing avenue for future study is to review the hyperparameter tuning by adding stratified sampling before cross-validation. Future study paths might include the National Research Council (NRC) of Canada's emotion lexicons, which include a broad range of attitudes such as Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, and Trust for sentiment analysis and categorization on microblogging sites. Furthermore, detecting disinformation about the COVID-19 pandemic is necessary to limit its spread. Finally, to categorize the COVID-19 tweets, pretrained transfer learning (TL) models such as bidirectional encoder representations from transformers (BERT) and a robustly optimized BERT pretraining technique (RoBERTa) may be used. We should also mention that this research concentrated on the diagnostic examination of society emotions. Related social media studies, for example, have tried to examine the purposeful manipulation of society emotions. An important future research path is to investigate the interaction between sentiment analysis and sentiment manipulation, for example, to detect purposeful attempts to sway public attitudes in certain ways.

REFERENCES

[1] R. Khan, R. Khan, P. Shrivastava, A. Kapoor, A. Tiwari, and A. Mittal, “Social media analysis with AI: Sentiment analysis techniques for the analysis of Twitter COVID-19 data,” J. Crit. Rev., vol. 7, no. 9, pp. 2761–2774, 2020.

[2] N. K. Rajput, B. A. Grover, and V. K. Rathi, “Word frequency and sentiment analysis of Twitter messages during Coronavirus pandemic,” Apr. 2020. [Online]. Available: arXiv:2004.03925.

[3] A. Kruspe, M. Häberle, I. Kuhn, and X. X. Zhu, “Cross-language sentiment analysis of European Twitter messages duringthe COVID-19 pandemic,” Aug. 2020. [Online]. Available: http://arxiv.org/abs/2008.12172.

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

[4] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the OMG!," in Proc. Int. AAAI Conf., 2011, pp. 538–541.

[5] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in Proc. Workshop Lang. Social Media, 2011, pp. 30–38.

[6] A. S. Imran, S. M. Doudpota, Z. Kastrati, and R. Bhatra, "Crosscultural polarity and emotion detection using sentiment analysis and deep learning—A case study on COVID-19," IEEE Access, vol. 8, pp. 181074–181090, 2020, doi: 10.1109/ACCESS.2020.3027350.

[7] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks," Expert Syst. Appl., vol. 164, Feb. 2021, Art. no. 114006. [Online]. Available: https://doi.org/10.1016/j.eswa.2020.114006

[8] J. Xue et al., "Twitter discussions and emotions about the COVID19 pandemic: Machine learning approach," J. Med. Internet Res., vol. 22, no. 11, Nov. 2020, Art. no. e20550, doi: 10.2196/20550.

[9] R. Abbas and K. Michael, "COVID-19 contact trace app deployments: Learnings from Australia and Singapore," IEEE Consum. Electron. Mag., vol. 9, no. 5, pp. 65–70, Sep. 2020, doi: 10.1109/MCE.2020.3002490.

[10] J. Zhou, S. Yang, C. Xiao, and F. Chen, "Examination of community sentiment dynamics due to COVID-19 pandemic: A case study from Australia," Jun. 2020. [Online]. Available: http://arxiv.org/abs/2006.12185.